Scientific Journal of Informatics
Vol. 8, No. 1, May 2021
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Object Detection in Last Decade - A Survey*

Usama Arshad[1*], Ali Raza[2]
[1,2]Department of Computer Science, COMSATS University Islamabad, Pakistan

Abstract
Purpose: In the last decade, object detection is one of the interesting topics that played an important role in revolutionizing the present era. Especially when it comes to computer vision, object detection is a challenging and most fundamental problem. Researchers in the last decade enhanced object detection and made many advanced discoveries using technological advancements.
Methods: This research work describes the advancements in object detection over the last 10 years (2010-2020). Different papers published in last 10 years related to object detection and its types are discussed with respect to their role in the advancement of object detection.
Result: This research work also describes different types of object detection, which include text detection, face detection etc. It clearly describes the changes in object detection techniques over the period of last 10 years. Object detection is divided into two groups. General detection and Task-based detection. General detection is discussed chronologically and with its different variants while task-based detection includes many state-of-the-art algorithms and techniques according to tasks. This paper also described the basic comparison of how some algorithms and techniques have been updated and played a major role in advancements of different fields related to object detection.
Novelty: This research concludes that the most important advancements that happened in last decade and future is promising much more advancement in object detection on the basis of work done in this decade.

Keywords: Computer Vision, Object Detection, Text Detection, Face Detection, YOLO, RCNN, Fast RCNN.
Received February 2021 / **Revised** February 2021 / **Accepted** March 2021

INTRODUCTION

Object detection is the detection of different kinds of objects and assigning them to classes; it is one of the fundamental tasks of computer vision [1]. Computer vision is an emerging field and its advancement is necessary for the overall revolution in computer science. Object detection involves some main elements like datasets, algorithms, and techniques. In object detection, normally have a dataset of some objects and this study wants the computer to detect those objects without telling the computer about objects. The whole process is divided into many parts and learning can be of many types. The types of learning include supervised and unsupervised learning. In supervised learning, the model labeled data to learn and then detect data in the testing phase while in unsupervised models no labeled data is provided to model and model learn by itself by calculation of loss [2]. As object detection is fundamental for computer vision, it is the basis of different tasks in computer vision. Some of the most important object detection based tasks include activity recognition [3], image captioning [4], face detection and recognition [5], object tracking, image segmentation, etc. According to its applications object detection can be divided into two types. One is general detection, which includes detection of real-world objects. Much work is being done on this especially using You ONLY LOOK ONCE (YOLO) algorithm. It real-time object detection [6]. The other type can be called application-based object detection, in this type object detection is used according to its need to get work done. The main work about object detection actually started after 2012 with the proposed models of Regional Convolutional Neural Network (RCNN) [7], FAST Regional Convolutional Neural Network (Fast RCNN) [8], and FASTER Regional Convolutional Neural Network (Faster RCNN) [9]. These models completely changed the working of historical models and brought a revolution in general object detection. Soon after them YOLO [10] and its three versions one after another made high contributions to advancement in object detection and different fields. However, yet study have to face many challenges in object detection to make it more fast and accurate.

For a long time, object detection faced issues related to speed, accuracy, and space. The models after 2012 were mostly two stage detectors [11], In the last decade different two stage and one stage models are

* Corresponding author.
Email addresses: usamajanjua9@gmail.com (Arshad), razaali4939@gmail.com (Raza)
DOI:10.15294/sji.v8i1.28956

proposed. The main issue of the models is that the single stage models are fast and simple however, they are less accurate as compared to two stage models. The main research work in single stage detectors [12] is done to increase the accuracy on different datasets, while the work done on two stage detectors is to increase the speed and simplicity of models. In the present era, object detection is being used on a vast scale due to its advancements. Real-world uses also include video surveillance for security purposes. The latest algorithms can easily detect any face in a crowd with high accuracy and precision making it easier for security agencies to find the criminals. Moreover, the industrial use is tremendous, Automatic robots are working in different industries doing the work faster and better. The Medical field is also being benefited from object detection. Especially after 5g technology, object detection can have much more efficiency than before. The last 10 years, object detection was not as good as it is today. Moreover, especially after 5g [13], cannot imagine what great future advancements that will have in the next 10 years. In this paper describes the two types of advancements in object detection. One is the advancement in general object detection and the second is the advancement in application-based object detection. The paper thoroughly describes and compare some of the most important techniques and models proposed and used in last 10 years. These methodologies played an important role in the advancement of computer vision and many other fields by making object detection better.

Datasets

The most important part of object detection is datasets. Many datasets are made in the last decade, which is publically available for researchers to use for their research to increase the effectiveness of object detection [14]. Some of most important and constantly used datasets include MS-COCO [15], Pascal VOC12 [16], ILSVRC [17], Open images [18]. Many researchers work hard to make datasets necessary for training and testing the latest algorithms and techniques for object detection. The available datasets are also constantly improved each year in yearly challenge events.

MS-COCO

It is one of the most used datasets since 2015. Since 2015 there is an annual competition held for improvement and innovations in the dataset. It has fewer object instances and more number object categories especially as compared to ILSVRC. As compared to other datasets commonly used this dataset is much better [15]. In MS-COCO per instance, segmentation is done and each object is labeled on instance level apart from bounding boxes. This kind of structure helps in better localization. Moreover, the dataset comprises more small and dense objects as compared to other datasets. With passing time, this dataset is widely used and has become a standard for researchers. Especially with YOLO algorithms, MS-COCO gives the best results. To explain the effectiveness used MS-COCO-17 as an example. This dataset contains 164k images, total of 80 categories, and 897k objects. This kind of dataset improves the overall process and makes object detection faster and accurate.

ILSVRC

ImageNet large-scale visual recognition challenge is another one of its kind events. It started in 2010 and after that, it was organized each year. The dataset produced and improved was used in many researches and is constantly used by researchers. The number of total visual objects present in the dataset are about 200. Images and objects may differ in different versions of datasets for example the ILSVRC-14 has around 517k images and number of objects used are around 534k [17].

Open Images

This dataset is more about task-oriented research rather than just visual detection. The dataset started in 2018 with Open Image Detection (OID) [18]. There are two main tasks for this dataset. The first task is standard object detection while the second task is the detection of visual relationships. This means more detection of relations between paired objects. The total images in the dataset are around 1910k. The total object category in the dataset is around 600 while the total bounding boxes are around 15440k.

Pascal VOC

Pascal VOC is one of the old challenges, which started in 2005 and remained till 2012. In this time period, the dataset performed well for different algorithms and tasks. The most used versions of Pascal VOC are VOC07 and VOC12 [16]. Different versions consisted of images from 5k to 12k. As many new and improved datasets were introduced like MS-COCO. The use of Pascal VOC became less with time. Now it is used as testing for some scenarios.

Specific Datasets

Apart from these datasets, that have many other task-based datasets used by researchers. Researchers may use many datasets that are already available or may introduce their own datasets for their research. Introduction of new dataset is also a great contribution and may help many other researchers in their research. Specifically, in task-based detection, researchers need specific datasets. For example, for pedestrian detection specific datasets are used only for the specific task.

METHODS

Two types of advancements in object detection are discussed. One is the advancement in general object detection and the second is the advancement in application-based object detection. The paper thoroughly describes and compares some of the most important techniques and models proposed and used in the last 10 years. These methodologies played an important role in the advancement of computer vision and many other fields by making object detection better. The best papers are taken depending upon their use and popularity in literature. Figure 1 clearly shows our methodology to collect the best research in chronological order.

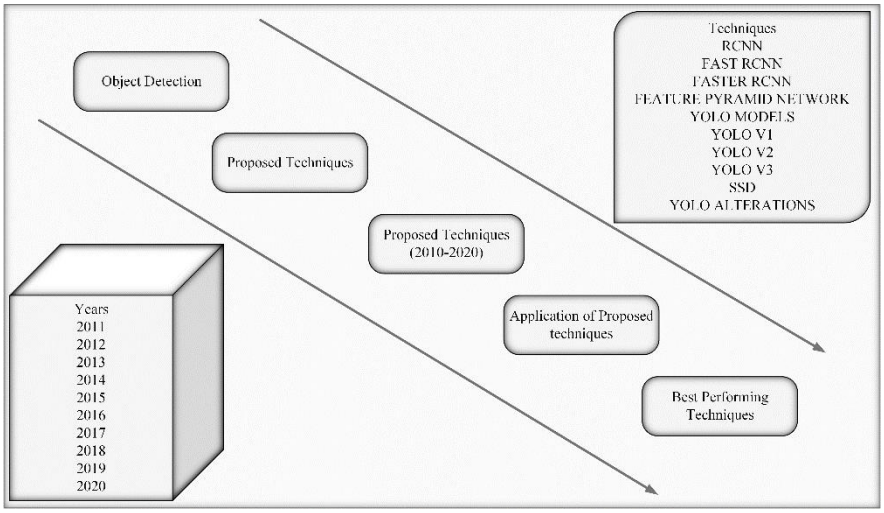

Figure 1. Methodology to Collect the Best Research in Chronological Order

General Detection

The work done on object detection often refers to a particular application or task. However, some algorithms and techniques are only focused on improving general object detection. In this type, choose the algorithm and weights and customize most of the things. Results may vary according to customizations. Machine learning is flourishing in sense of computer vision especially object detection and fast-moving object detection [19]. Mostly deep convolution neural networks [20] are used for this purpose trained on Image Net dataset [21]. Faster RCNN and YOLO are currently used. Old methods of object detection are for low-resolution images and need more methods for high-resolution images and videos. Low-resolution images are mostly used because they are time and cost efficient in most cases but they sacrifice accuracy. The data sets publically available are also mostly comprised of low-resolution images. With the latest cameras, using high resolution data so need models to process these data for object detection. Current models are limited in terms of high-resolution data, they use two base approaches either to downscale an image to detect objects which sacrifice accuracy and most of data is lost or cropping image into smaller parts to detect objects but this is costly in terms of speed [22]. Talk about the last decade, improved object detection slowed down from 2010 to 2012. However, era after 2012 shows great improvement in object detection models. Especially from 2014 onwards with the model of RCNN proposed. These models gave new life to convolutional neural networks. However, The models after 2012 were mostly two stage detectors. In the last decade, different two stage and one stage models are proposed. The main issue of the models is that the single stage models are fast and simple however, they are less accurate as compared to two stage models. The main research work in single stage detectors is done to increase the accuracy on different datasets, while the work done on two stage detectors is to increase the speed and simplicity of models.

RCNN

Region based Convolutional Neural Network (RCNN) played an important role in object detection from a long time [7]. The main idea behind RCNN is to select some parts of image and detect. Now the next question was how the area is selected and how much region is selected. The image can consist of huge area that do not need to process all regions to solve this process only 2k region is used called as proposal. This proposal is selected by a selection algorithm. The selected parts are converted to fixed size images and then sent to a convolution network for detection. This convolutional neural network is normally trained on a dataset of ImageNet. This made things easier for object detection for a long time but as it had many drawbacks like those that have to select 2k regions from a single image making detection slower. When you look at the structure of RCNN given in Figure 2 taken from [7] can easily see four different parts of structure.

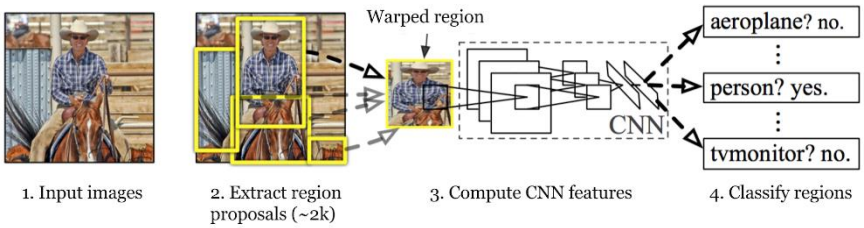

Figure 2. RCNN Structure [7]

In first part input image is taken by the structure. In second part Region proposals are extracted from image and these proposals consist of 2k regions. The the model computes the convolutional neural network (CNN) features. In part, four regions are classified. However soon after RCNN in 2014, Spatial Pyramid Pooling Networks (SPPNET) [23] was proposed which solved many such problems. The best thing done in this model was that the model only looks once on the image to get regions, Secondly, as the RCNN it does not have to take fixed image sizes, It can work on image of any size and generate fixed size representations. SPPNET solved the issue of speed, however it was still not good in some aspects. For example, it was performing best for fully connected layers but not the same for partially connected layers. In 2015, these problems were solved as the Fast RCNN was proposed. Much work was done using SPPNET for this period of time. Some of the tasks done using SPPNET include image processing, Food classification [24], Semantic Segmentation [25].

Fast RCNN

The author, who proposed RCNN, also proposed its modified version and called it Fast RCNN. Fast RCNN solved almost all the existing problems it had all the positive aspects of RCNN and SPPNET. It was faster and the accuracy was better than old versions. The reason it got better was it was now not taking 2k regions again and again, it simple takes the image once and generates an image map from it. From generated image map, image proposals were taken. Now the main issue was without the proposal Fast RCNN was very fast however with proposals, it was a bit slow. To solve this issue author shortly proposed Faster RCNN. The Figure 3 briefly describes the structure of Fast RCNN taken from [8].

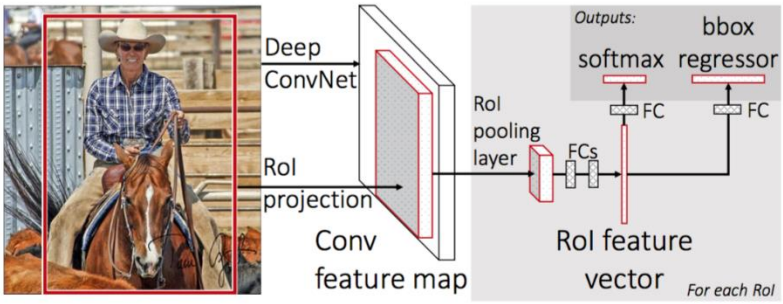

Figure 3. Fast RCNN Structure [8]

Faster RCNN

Faster RCNN solved the existing problems and increased the speed and efficiency. This model was different from last approaches, In RCNN and Fast RCNN, a selective algorithm was used, which was really time

consuming. Selective algorithm was used to look for regions. However, in this new model of Fast RCNN, another network did the process of selection. Just like old system, image was taken, image map is generated from image but instead of using selective algorithm another network selects the regions and other processes remain same. This approach was fast and accurate and can even be used for real time object detection. The Figure 4 briefly describes the structure of Faster RCNN taken from [9]. Many other models like Mask RCNN [26] were also proposed in this time period.

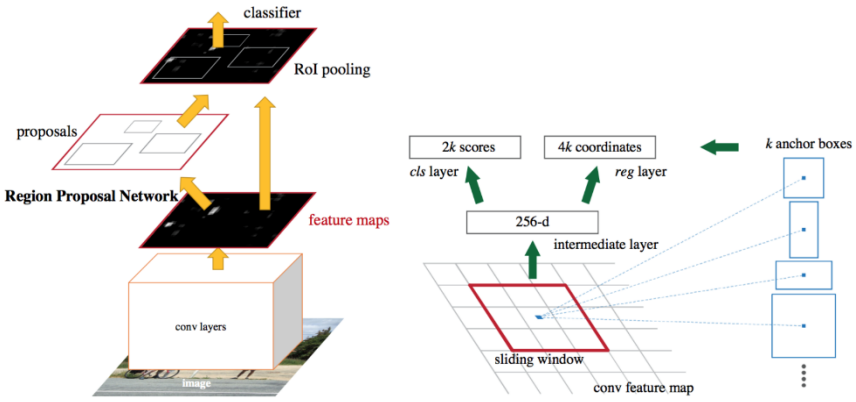

Figure 4. Faster RCNN Structure [9]

Feature pyramid Networks

This model was proposed in 2017 and works on the model of Faster RCNN. The issue before this model was most of the work was done on the top layer but after this model deep detection was done on deep layers and this also made easier to recognize different categories. Moreover, Convolutional Neural Network (CNN) already works on the basis of layers in forward pass so using Feature pyramid Networks (FPN) provides the advantage. After the Faster RCNN, FPN became the base of many new detectors and models. Since then FPN is used in different detection models with slight changes for improvement [27]. Figure 5 briefly describes the structure of the Feature pyramid Network taken from [28]. Many single shot detectors are also used based on FPN for some time [29]. Deep feature pyramid networks were also used for object detection [30].

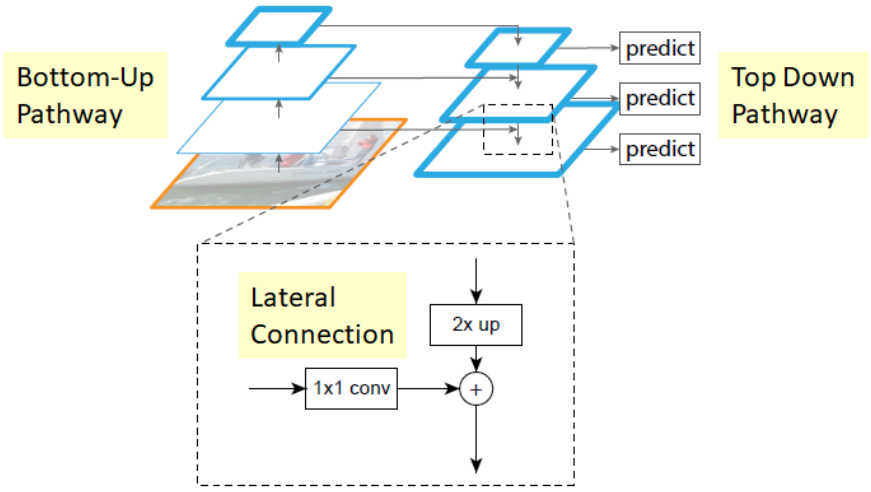

Figure 5. Feature pyramid Network Structure

YOLO MODELS

YOLO stands for you only look once and that is the main concept behind it. Idea is to make this system so perfect that it can detect the objects in just one go. YOLO is a whole system for object detection that is used

in different researches for few last years to accomplish tasks. YOLO is customized as needed to get the best results possible. YOLO came out in 2016 and is used for different researches. After the YOLO it also got updated to YOLO-v2 and YOLO-v3. YOLO is constantly used in general object detection with customization. Paper [31] provides techniques for fast and accurate object detection. High-resolution data is downscaled and divided into parts for detection. Only specific areas in original image are focused on. This saves data loss and helps in efficient data detection. Many other Yolo models are proposed for efficient object detection by modification in YOLO. [32] Proposed a Yolo-lite model to be used on portable devices without any need of GPU. The goal is to get the accuracy of 30 using data sets of PASCAL VOC. Model is first trained on PASCAL VOC and then on coco datasets to achieve maximum accuracy and a table is provided which shows the comparison between different Yolo models.

YOLO v1

Yolo is fast in case of object detection. The detection speed in real-time goes up to 45 fps in normal cases. Speed of 155 fps can be achieved with fast Yolo. Yolo only looks once and detects objects. It creates S x S (7 x 7) grid on input and creates bounding boxes and does call probability mapping to check probability of detecting of an object in grid cells. Every grid cell finds bounding boxes (b where b=2) and confidence for those boxes. Bounding boxes are generated and they consist of five predictions including x-axis, y-axis, height, width, and confidence scores. Class probabilities are also found to detect the probability of detecting the object. According to the structure model, Yolo model consists of 24 convolutional layers and two fully connected layers.

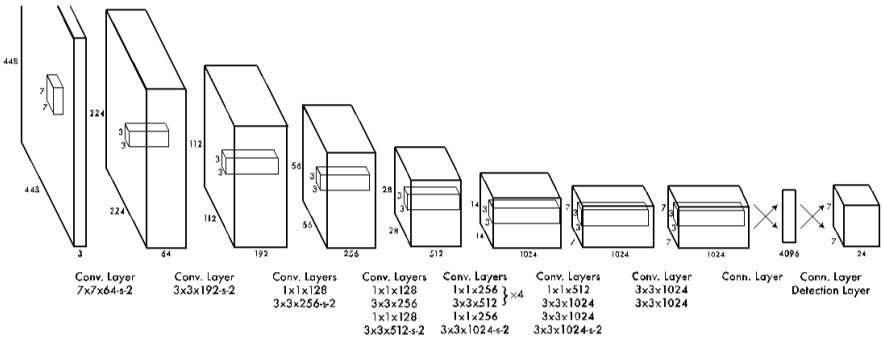

Figure 6. Yolo v1 Structure [10].

Different other models are proposed after modification in Yolo. Fast Yolo consists of less number of convolution layers (9). Input image only passes once from the network and the object is detected. Figure 6 briefly describes the structure of Yolo v1 taken from [10]. Loss function clearly describes different parts of algorithms and how they are contributing towards the betterment of algorithm as shown in Figure 7 below.

$$
\begin{aligned}
&\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
&+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \qquad (3)
\end{aligned}
$$

Figure 7. Loss Function (yolov1).

The activation function used in Yolo v1 is leaky relu.

YOLO v2

Yolov2 is improved version of Yolo; the main goal was to increase accuracy with speed [33]. Bounding boxes are more diverse and accuracy is higher. In yolo v1 had less bounding boxes and boxes were guessed arbitrarily. According to the structure, fully connected layers are removed. Predictions are moved from cell level to boundary level. Yolo v2 outperformed Yolo v1 with the new improvements and produce better results than many other models. The Figure 8 briefly describes the structure of Yolo v2 taken from [34].

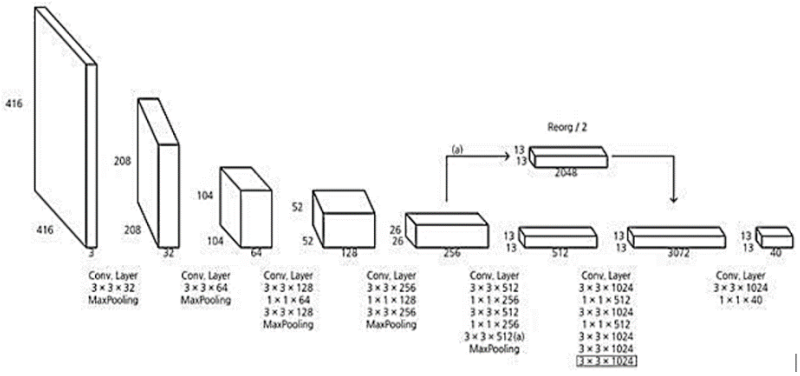

Figure 8. Yolo v2 Structure [33].

This model was later improved in Yolo v3. The structure clearly describes all the layers. The convolutional layers are shown with the max-pooling and complete information about the filter and outputs.

YOLO v3

Better than Yolo v2 on accuracy but slower due to complex, structure. The main difference between yolov2 and yolov3 is that yolov3 uses logistic functions to find an object, unlike yolov2 where probability is checked [34]. Figure 9 briefly describes the structure of Yolo v3 taken from [34].

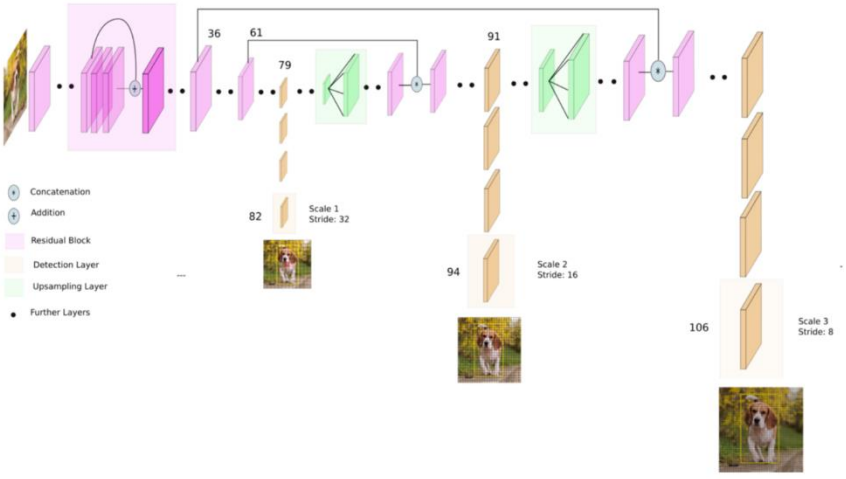

Figure 9. Yolo v3 Structure.

SSD

After different versions of Yolo, Single Shot Multibox Detector (SSD) was proposed [35]. This model is used as an alternative with Yolo for a long time. SSD is good as compared to basic Yolo v1 but when it comes to Yolo v3, SSD sometimes does not perform as well as Yolo v3. Like other models, SSD also passes input images through different layers of convolutional neural network. Different bounding boxes are generated on different scales. Hence results in better accuracy and performance. Figure 10 demonstrates the difference between structures of Yolo and SSD taken from [35]. The structures clearly describe the concept of fully connected and partially connected layers. SSD challenges Yolo and outperforms Yolo.

After the SSD many other detection models and techniques like Ratina Net [36] are proposed which are used sometimes by researchers for specific tasks. However, for general detection, Yolo and SSD are still used on a large scale. Many variants of SDD with different customization are also used. Feature Fusion Single Shot Detection (FSSD) is also used by many researchers [37]. Many other variants were proposed for fast detection on different datasets [38]. There are many future object detection advancements with combined models of SSD and other Yolo models with two-stage detectors.

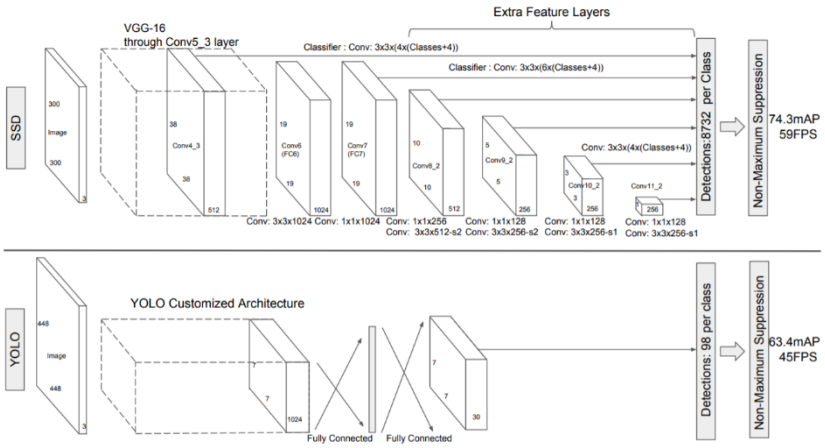

Figure 10. Difference between YOLO & SSD

Yolo alterations

In computer vision, object detection is a fundamental challenge. Object detection is used in every field like image processing, pattern recognition, machine learning, etc. Authors in paper [33], proposed an attention pipeline. In this attention pipeline, two things are covered. One is to downscale images and 2nd is cropping the real image into parts and matching both images in a way to only check the focused parts of image termed this concept as attention. The model first downscales the original image and then in low-resolution image only focuses on specific parts after looking at the real image and highlighting objects. Further research based on a paper [33], in which a YOLO lite model is proposed. [32][41] A YOLO lite model is proposed that is used on portable devices without any need for GPU. The goal is to get the accuracy of 30 using data sets of PASCAL VOC. Model is first trained on PASCAL VOC and then on coco datasets to achieve maximum accuracy and a table is provided which shows the comparison. Hun Yang and Rui Li proposed an improved YOLOv2 in their paper [39][42]. The proposed approach solves the problem of large parameters of convolution layers. This paper introduced a Depth-wise Separable Convolution. It reduces the number of parameters and enhances the speed of YOLOv2 detection model. The problem in detection of small object is resolved through this paper.

YOLOv2 uses a feature fusion method that weakens the performance of YOLOv2. In this proposed improved YOLOv2, the feature fusion method is replaced with Feature Pyramid Network. It can detect objects with different sizes to improve detector precision. The recognition accuracy became sensitive to detect smaller objects. It has higher accuracy for detecting small objects. An improved YOLOv2 is faster and detects smaller objects. The parameters of convolution layers are reduced by 78.83%. However, it is more complex than YOLOv2. It is complex to detect smaller objects, especially in complex scenarios. An improved structure is proposed [40][43], in which smaller objects are detected and accurately recognized. It reduced the complexity of the YOLOv2 model. To improve the detection accuracy 1 x 1 convolution layer is proposed and the output size is changed to extract more features from multi pixels images. To adapt the size of objects, the loss function is adapted in this model. Paper [41] proposed an object detection method based on lightweight network architecture. The proposed model consists of a detector, which consists of a single-stage network model based on YOLOv2 architecture. By reducing number of channels in each convolutional layer makes it lightweight. This network is also embedded with extended path through layer, which helps to detect smaller objects more easily. In this paper [42], a lightweight YOLOv2 is proposed which consists of binarized CNN and parallel support vector regression for object feature detection. For object feature detection, binarized CNN is used and parallel Support Vector Regression (SVR) is used for localization and classification. SVR is based on Support Vector Machine (SVM) [43],

which is a supervised learning model that is used for analyzing data for regression and classification analysis. Yolo with deep learning models is also used as needed with customizations [44].

Task Based Detection

Task-based object detection is grouped separately from general object detection but they are not much different from general detection. In fact, general object detection is used as the base, and with modifications; it is used for a specific task. The main examples of task-based detection include face detection, pedestrian detection, text detection, traffic signals, and traffic light detection. Task-based detection takes the modern algorithms and techniques and uses them with new focused approach to accomplish a task with maximum accuracy.

Experiments

The experiments are taken on Intel Core i5 7200U CPU, 2.50 GHz, and Windows 10. The dataset is COCO dataset. It is an excellent large dataset designed to detect objects with 120,000 images, in which 80,000 images are training images and the rest images are validation images. The table shows the models and their accuracy against different objects. Darknet repository is used to implement YOLO. Darknet is a framework of Neural Network (NN) written in CUDA and C language.

RESULT AND DISCUSSION

The results show that the YOLOv3 model performs better than other object detection models. SSD sometimes performs well and sometimes not performs as well as other models are performing. SSD challenges YOLO. However, YOLOv1 is better than SSD but this model is slow. The mean Average Precision (mAP) of SSD is 41.6% whereas the mAP of YOLOv1 model is 58.1%. YOLOv2 is a better version of YOLOv1. It predicts objects accurately and faster than YOLOv1. The mAP of YOLOv2 is 62.5%. YOLOv3 is a better version of the object detection model than other versions of YOLO. It always predicts objects accurately above 50%. The mAP of YOLOv3 model is 82%. However, this model is more complex and takes more computational powers and resources than other detection models. Figure 11 briefly describes comparison of models.

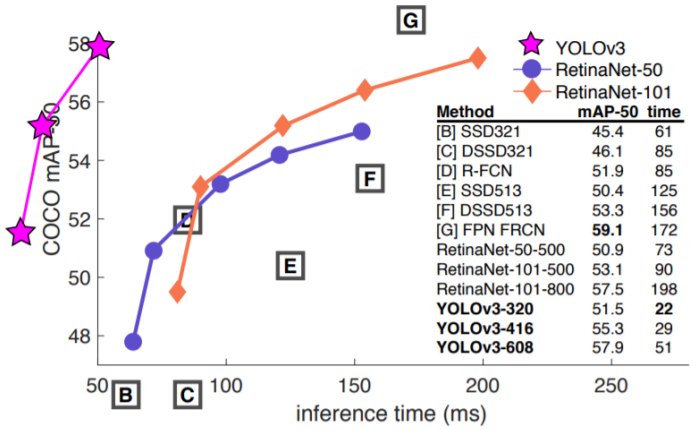

Figure 11. Comparison of Models

Table 1 and Table 2 clearly describe the accuracy and finishing time for object detection. This comparison of models clearly shows the difference in results provided by each model for different objects.

Table 1. Accuracy Comparison of Models

Models	Bicycle	Dog	Cellphone	Cat	Bottle	Bird	Stop-sign	Person	Laptop	Chair
SSD	0.47	0.51	0.45	0.35	0.30	0.32	0.52	0.45	0.48	0.36
YOLOv1	0.93	0.38	0.35	0.82	0.40	0.70	0.37	0.67	0.35	0.98
YOLOv2	0.68	0.64	0.70	0.64	0.34	0.75	0.54	0.63	0.71	0.71
YOLOv3	0.89	0.97	0.72	0.81	0.73	0.98	0.79	0.99	0.73	0.79

Table 2. Finishing time taken by Models

Models	Bicycle	Dog	Cellphone	Cat	Bottle	Bird	Stop-sign	Person	Laptop	Chair
SSD	0.121	0.121	0.143	0.125	0.171	0.121	0.121	0.181	0.121	0.131
YOLOv1	0.13	0.11	0.13	0.15	0.13	0.1	0.11	0.11	0.101	0.1
YOLOv2	0.124	0.112	0.125	0.134	0.143	0.114	0.122	0.129	0.112	0.114
YOLOv3	1.25	1.29	1.251	1.251	1.22	1.22	1.513	1.175	1.176	1.345

CONCLUSION

Object detection is advancing on a rapid level especially after 2012 and it holds much potential after 2020. All the models from 2012 to 2020 made object detection faster and accurate and there is no stopping in the improved accuracy and speed. New datasets are made for specific tasks to detect better and faster. Especially after 5g future holds much potential for improvement in different models in terms of speed and accuracy. Real-time object detection is facing most of the issues in 2020. However, after 5g the speed and accuracy of models may increase by 5 to 10 folds. This study have many issues related to task-based detections as researchers are always focused on general object detection rather than a specific task. However, as the general object detection improved, task-based detection also improves accordingly. Different fields like Computer vision already improved much after advancements of object detection and as it the object detection improves more these fields will have a promising future. Task-based object detection has so much potential when it comes to real-life applications. Many new models are proposed especially in the last 3 to 5 years and this paper may see different new models joined with old models to achieve high accuracy and speed in the future.

REFERENCES

[1] R. Szeliski, *Computer vision: algorithms and applications*. Springer, 2010.

[2] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *Int. J. Adv. Res. Artif. Intell.*, vol. 2, no. 2, pp. 34–38, 2013.

[3] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 790–808, 2012.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4651–4659, 2016.

[5] F. Ahmad, A. Najam, and Z. Ahmed, "Image-based Face Detection and Recognition: 'State of the Art,'" *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 3–6, 2013.

[6] M. Mehta, C. Goyal, M. C. Srivastava, and R. C. Jain, "Real time object detection and tracking: Histogram matching and Kalman filter approach," *2010 2nd Int. Conf. Comput. Autom. Eng. ICCAE 2010*, vol. 5, pp. 796–801, 2010.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 580–587, 2014.

[8] R. Girshick, "Fast R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1440–1448, 2015.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 779–788, 2016.

[11] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: In defense of two-stage object detector," *arXiv*, pp. 1–9, 2017.

[12] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single Stage Headless Face Detector," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 4885–4894, 2017.

[13] J. G. Andrews *et al.*, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[14] N. Dvornik, J. Mairal, and C. Schmid, "Modeling visual context is key to augmenting object detection datasets," *Proc. Eur. Conf. Comput. Vis.*, pp. 364–380, 2018.

[15] T. Yi-Lin *et al.*, "Microsoft COCO," *Eur. Conf. Comput. Vis.*, pp. 740–755, 2014.

[16] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild," *IEEE Winter Conf. Appl. Comput. Vis.*, pp. 75–82, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks Alex," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[18] A. Kuznetsova *et al.*, "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, 2020.

[19] K. Choi and J. Yun, "Robust and Fast Moving Object Detection in A Non-Stationary Camera Via Foreground Probability Based Sampling," *IEEE Int. Conf. Image Process.*, pp. 4897–4901, 2015.

[20] H. J. Yoo, "Deep Convolution Neural Networks in Computer Vision: a Review," *IEIE Trans. Smart Process. Comput.*, vol. 4, no. 1, pp. 35–43, 2015.

[21] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[22] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Adv. Neural Inf. Process. Syst.*, pp. 379–387, 2016.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.
[24] E. Jahani Heravi, H. Habibi Aghdam, and D. Puig, "An optimized convolutional neural network with bottleneck and spatial pyramid pooling layers for classification of foods," *Pattern Recognit. Lett.*, vol. 105, pp. 50–58, 2018.
[25] H. Li, P. Xiong, J. An, and L. Wang, "Dynamic attention network for semantic segmentation," *Neurocomputing*, vol. 384, pp. 182–191, 2018.
[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020.
[27] W. S. Lai, J. Bin Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 5835–5843, 2017.
[28] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 936–944, 2017.
[29] Q. Zhao *et al.*, "M2Det: A single-shot object detector based on multi-level feature pyramid network," *Thirty-Third AAAI Conf. Artif. Intell.*, 2018.
[30] T. Kong, F. Sun, W. Huang, and H. Liu, "Deep Feature Pyramid Reconfiguration for Object Detection," *Eur. Conf. Comput. Vis.*, vol. 1, pp. 172–188, 2018.
[31] V. Ruzicka and F. Franchetti, "Fast and accurate object detection in high resolution 4K and 8K video using GPUs," *IEEE High Perform. Extrem. Comput. Conf.*, 2018.
[32] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: A Real-Time object detection algorithm optimized for non-GPU computers," *IEEE Int. Conf. Big Data*, pp. 2503–2510, 2018.
[33] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6517–6525, 2017.
[34] J. Redmon and A. Farhadi, "YOLO v.3," *arXiv:1804.02767*, pp. 1–6, 2018.
[35] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, "SSD: Single Shot MultiBox Detector," *Eur. Conf. Comput. Vis.*, vol. 1, pp. 21–37, 2016.
[36] W. Khan, N. Zaki, and L. Ali, "Intelligent Pneumonia Identification from Chest X-Rays: A Systematic Literature Review," *medRxiv*, 2020.
[37] Z. X. Li and F. Q. Zhou, "FSSD: Feature fusion single shot multibox detector," *arXiv:1712.00960*, 2017.
[38] L. Zheng, C. Fu, and Y. Zhao, "Extend the shallow part of single shot MultiBox detector via convolutional neural network," *arXiv:1801.05918*, 2018.
[39] R. Li and J. Yang, "Improved YOLOv2 Object Detection Model," *Int. Conf. Multimed. Comput. Syst. - Proceedings*, vol. 2018-May, pp. 1–6, 2018.
[40] E. Dong, Y. Zhu, Y. Ji, and S. Du, "An improved convolution neural network for object detection using Yolov2," *IEEE Int. Conf. Mechatronics Autom.*, pp. 1184–1188, 2018.
[41] Y. C. Lim and M. Kang, "Object Detection Using a Single Extended Feature Map," *IEEE Intell. Veh. Symp. Proc.*, vol. 2018-June, no. Iv, pp. 820–825, 2018.
[42] H. Nakahara, Y. Haruyoshi, T. Fujii, and S. Sato, "A Lightweight YOLOv2: A Binarized CNN with A Parallel Support Vector Regression for an FPGA," in *Proc. 2018 ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, 2018, pp. 31–40.
[43] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.
[44] W. Hammedi, M. Ramirez-Martinez, P. Brunet, and S.-M. Senouci, "Deep Learning-Based Real-Time Object Detection in Inland Navigation," *IEEE Glob. Commun. Conf.*, pp. 1–6, 2019.